DOCUMENT RESUME

ED 074 133 TM 002 496

AUTHOR Moy, Mabel L. Y.; Barcikowski, Robert S.
TITLE Item Sampling: Optimum Number of People and Items.
PUB DATE Feb 73
NOTE 28p.; Paper presented at annual meeting of the National Council on Measurement in Education, AERA (New Orleans, La., February 25-March 1, 1973)

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Evaluation Techniques; *Item Sampling; *Sampling; Speeches; *Standard Error of Measurement; Statistical Studies; Tables (Data); *Tests
IDENTIFIERS Monte Carlo Methods

ABSTRACT

Using a computer-based Monte Carlo approach to generate item responses, the results of this study indicate that, when item discrimination indices are considered, item-examinee sampling procedures having the same number of observations have different standard errors in estimating both test mean and test variance. With certain types of tests, a single item-examinee sampling plan would not yield optimal, i.e., smallest standard error, estimates of both mu and sigma squared. That is, one sampling plan would be needed to optimally estimate mu and another to optimally estimate sigma squared. In addition, it was found that single exhaustion of the item set was sufficient for estimating both mu and sigma squared. (Author)

FILMED FROM BEST AVAILABLE COPY

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

ED 074133

TM 002 496

ITEM SAMPLING: OPTIMUM NUMBER OF PEOPLE AND ITEMS

Mabel L. Y. Moy

and

Robert S. Barcikowski
Ohio University

A Paper Presented at the Annual Meeting of the National Council on Measurement in Education, February, 1973.

ABSTRACT

Using a computer-based Monte Carlo approach to generate item responses, the results of this study indicate that, when item discrimination indices are considered, item-examinee sampling procedures having the same number of observations have different standard errors in estimating both test mean and test variance. With certain types of tests, a single item-examinee sampling plan would not yield optimal, i.e., smallest standard error, estimates of both μ and σ^2. That is, one sampling plan would be needed to optimally estimate μ and another to optimally estimate σ^2. In addition, it was found that single exhaustion of the item set was sufficient for estimating both μ and σ^2.

INTRODUCTION

With the need for many and continuous evaluation studies to be performed in the service of improved instruction in our schools, and considering the fact that there are always limitations of time, money and personnel with which to perform such evaluations, it is important that procedures be used which not only provide accurate information, but are economical as well. Item sampling has been suggested as just such a procedure. With item sampling, the savings, particularly in test-taking time, can be enormous.

However, when faced with an evaluation project, how should school personnel proceed in implementing an item sampling procedure? What guidelines are there concerning the optimum number of items and examinees to use in a particular situation? These questions could be answered if information were available on the standard error in estimating a test's mean and variance under conditions similar to that to be encountered in this project. Ideally, a sampling plan would be chosen that would yield a relatively small standard error of the mean and/or variance.

Barcikowski (1970) and Shoemaker(1971), have indicated that the particular sampling plan chosen does make a difference. Barcikowski, in particular, called attention to the need to consider the range of biserial correlations between item response and ability, i.e., item discrimination. With

reference to estimating the population mean by item sampling, both Barcikowski's and Shoemaker's findings support the use of a large number of subtests. Barcikowski's study, in addition, would recommend the use of small sized subtests. With reference to the estimation of the population variance, Barcikowski's findings contradicted those of Shoemaker. Whereas Shoemaker contended that item sampling plans having the same number of observations have for all practical purposes the same standard error in estimating the variance, Barcikowski found that this was not true when item biserial correlations were taken into consideration. In the case of tests with a range of biserial correlations between .40-.70, item sampling plans with small subtestsizes (e.g., 5 items) produced the best estimates of test variance. In the case of tests with a range of biserial correlations between .05 and .50, item sampling plans with subtests somewhat less than half the size of the whole test produced the best estimates of test variance. He concluded that optimal estimates of both the mean and variance from a single item sampling plan may not be possible.

Most item sampling studies have employed a sampling plan which might be described as "single exhaustion," in which all of the items on the whole test are used and each item appears on only one subtest. Barcikowski employed "multiple exhaustion," in which all the items on the whole

test are used, but the formation of subtests is continued by replacement of the set of items each time they are exhausted. There is a need to determine which of the two procedures is the more advantageous.

Specifically, this study was designed to provide answers to the following questions:

1. Given tests with various ranges of biserial correlations between item response and ability, i.e., item discrimination, what is the optimum number of items and examinees to be used with item-examinee sampling?

2. With item-examinee sampling, should single or multiple exhaustion of the set of items on a test be employed?

METHOD

The procedures to generate item responses for the tests used in this study are described by Barcikowski (1970). Briefly, the method used was Tucker's (1946) mathematical model of an item trace line. The item trace lines for examinees took into account item discrimination, item difficulty, and the ability range of the examinee, and were assumed to be of normal ogive form. The model allows direct computation of test population mean and variance.

Since differences in the range of item difficulty indices did not affect Barcikowski's results, only one range of rectangularly distributed item difficulties(.16 to .84, centered at .50) was used. This was chosen as appropriate

for a single factor test.

Table 1 presents the ranges of biserial correlations between item response and ability used to create seven whole tests. Each tests consisted of 60 items whose biserial correlations were randomly selected from the respective range, and whose item difficulties were randomly selected from the rectangularly distributed range of .16 to .84. A 60-item test was thought to be representative of a test from a standard achievement battery. A people sample size of 100 was chosen so that the probability of obtaining a sample test mean within .25 of the population mean would be .95. With a test of 60 items and a sample of 100 people, the total number of responses was fixed at 6000 (60 X 100).

The total number of responses was held constant at 6000, and the total test size was held constant at 60 items. The size of the subtests was varied over values of 6, 10, 15, 20, and 30 items. This study employed single exhaustion, multiple exhaustion, and an extreme situation in which the number of people taking each subtest was two. Table 2 presents the number of people that took each subtest and the number of subtests involved for each of the above-mentioned plans. The product of the number of subtests (t), the subtest size (k), and the number of people that took each subtest (n) equals the total of number of observations or responses (6000).

TABLE 1

RANGES OF BISERIAL CORRELATION BETWEEN ITEM RESPONSE AND ABILITY USED IN CONSTRUCTION OF WHOLE TESTS

	POSITION OF RANGE		
SIZE OF RANGE	LOW	MIDDLE	HIGH
.1	.10-.20	.50-.60	.80-.90
.5	.05-.55	.20-.70	.45-.95
.9		.05-.95	

TABLE 2

NUMBER OF ITEMS ON EACH SUBTEST AND THE CORRESPONDING NUMBER OF SUBTESTS AND NUMBER OF PEOPLE TAKING EACH SUBTEST UNDER SINGLE EXHAUSTION, MULTIPLE EXHAUSTION, AND EXTREME ITEM SAMPLING

	SINGLE		MULTIPLE		EXTREME	
SUBTEST SIZE	Number of Subtests	Number of People	Number of Subtests	Number of People	Number of Subtests	Number of People
6	10	100	20	50	500	2
10	6	100	20	30	300	2
15	4	100	20	20	200	2
20	3	100	20	15	150	2
30	2	100	20	10	50	2

One hundred estimates of test mean and variance were acquired for each subtest size. The population means and variances were then used to compute the sums of squared errors $\sum_{k=1}^{100} (\hat{\mu}_k - \mu_k)^2$, abbreviated SSEM, and $\sum_{k=1}^{100} (\hat{\sigma}_k^2 - \sigma_k^2)^2$, abbreviated SSEV.

Traditional sampling with 6000 total number of responses was studied by giving each of the seven whole tests to 100 people and obtaining 100 estimates of test mean and variance per whole test. The population means and variances were then used to compute the sums of squared errors $\sum_{T=1}^{100} (\hat{\mu}_T - \mu_K)^2$ and $\sum_{T=1}^{100} (\hat{\sigma}_T^2 - \sigma_K^2)^2$.

The estimates of test mean and variance were compared to the population means and variances and to each other by the use of these sums of squared errors. If, for example, the sum of squared errors were 5 for one sampling plan and 10 for another sampling plan, the ratio of SSE's would be less than one, thus indicating the superiority of the former over the latter sampling plan.

RESULTS AND DISCUSSION

The actual population means and variances of the seven whole tests used in this study are presented in Table 3. In this table the tests are listed according to the size of the range of biserial correlation, and within each size according to the size of the variance associated with them.

The results of estimating test population mean for the various sampling plans are given in Tables 4, 5, and 6. It can be seen that as total test variance decreased, and consequently the average value of the biserial correlations of the tests as well, estimates of the population mean improved. One explanation of this occurance is found in the relationship between the biserial correlations and test variance. As test biserial correlations decreased, test varaince also decreased, thus reducing the standard error of the mean.

The data presented in Tables 4, 5, and 6 also indicate that the better estimates of the mean are given by sampling plans with fewer items (smaller k's) and more subtests (larger t's). As the subtest size increased and the number of subtests decreased, the estimates became poorer, as evidence by the larger SSEM's. For example, in Table 4, tests with average biserial correlations of .80-.90 under the sampling plan of 10 subtests and 6 items per suntest (10/6/100) had SSEM of 43.63, whereas tests with the same average biserial correlations of .80-.90, but under the sampling plan of 2 subtests and 30 items

TABLE 3

PARAMATERS FOR WHOLE TESTS

Test Properties		Parameters	
Size of Range	Biserial Correlation	Mean	Variance
	.80-.90	29.91	326.87
.1	.50-.60	31.54	142.15
	.10-.20	31.72	23.05
	.45-.95	32.75	228.39
.5	.20-.70	29.50	115.18
	.05-.55	29.79	48.25
.9	.05-.95	29.53	135.66

TABLE 4

SUMS OF SQUARED ERRORS FOR MEANS UNDER TRADITIONAL AND SINGLE EXHAUSTION ITEM SAMPLING

Test Properties		Multiple Exhaustion Item Sampling Sums of Squared Errors					Traditional Sums of Squared Errors
Variance	Biserial Correlation	10/6/100	6/10/100	4/15/100 (t/k/n)[a]	3/20/100	2/30/100	1/60/100
326.87	.80-.90	43.63	42.65	70.35	100.64	184.18	328.70
228.89	.45-.95	42.13	41.60	76.71	75.96	103.57	231.91
142.15	.50-.60	27.91	21.54	42.70	62.79	83.25	1[illegible].73
135.66	.05-.95	21.66	34.97	37.75	53.58	72.30	165.25
115.18	.20-.70	20.90	28.73	38.68	48.44	56.05	92.58
48.28	.05-.55	19.34	14.82	19.61	27.35	30.77	43.81
23.05	.10-.20	12.06	15.12	13.83	14.38	21.83	25.57

[a]Code: t=number of subtests, k=number of items per subtest, n=number of examinees per subtest

TABLE 5

SUMS OF SQUARED ERRORS FOR MEANS UNDER TRADITIONAL AND MULTIPLE EXHAUSTION ITEM SAMPLING

Test Properties		Multiple Exhaustion Item Sampling Sums of Squared Errors					Traditional Sums of Squared Errors
Variance	Biserial Correlation	20/6/50	20/10/30	20/15/20 (t/k/n)[a]	20/20/15	20/30/10	1/60/100
326.87	.80-.90	30.49	65.41	91.04	103.75	128.00	328.70
228.89	.45-.95	27.98	50.24	59.83	64.01	127.77	231.91
142.15	.50-.60	21.14	43.56	47.44	57.18	68.10	179.75
135.66	.05-.95	25.44	36.67	51.95	61.24	77.61	165.25
115.18	.20-.70	24.59	28.10	46.91	48.91	59.68	92.58
48.28	.05-.55	15.10	22.84	20.39	20.71	33.33	43.81
23.05	.10-.20	12.54	16.81	15.58	19.78	22.15	25.57

[a]Code: t=number of subtests, k=number of items per subtest, n=number of examinees per subtest

TABLE 6

SUMS OF SQUARED ERRORS FOR MEANS UNDER TRADITIONAL AND EXTREME ITEM SAMPLING

Test Properties		Extreme Item Sampling Sums of Squared Errors					Traditional Sums of Squared Errors
Variance	Biserial Correlation	500/6/2	300/10/2	200/15/2 (t/k/n)[a]	150/20/2	100/30/2	1/60/100
326.87	.80-.90	30.93	45.57	89.14	122.60	175.61	328.70
228.89	.45-.95	24.39	41.37	48.57	87.88	123.43	231.91
142.15	.50-.60	18.08	30.37	56.45	51.60	78.73	179.73
135.66	.05-.95	22.00	35.84	38.71	55.59	73.13	165.25
115.18	.20-.70	17.91	28.34	36.06	47.87	51.72	92.58
48.28	.05-.55	13.69	17.84	27.22	22.94	27.53	43.81
23.05	.10-.20	12.38	15.03	11.99	14.76	16.60	25.57

[a]Code: t=number of subtests, k=number of items per subtest, n=number of examinees per subtest

per subtest (2/30/100), had a SSEM of 184.18. Traditional sampling (k=60, t=1) supplied the poorest estimate, with a SSEM of 328.70. These results agree with the theory presented in Lord and Novick (1968), and served as a partial check on the model.

The data in Tables 7, 8, and 9 present comparisons of single exhaustion, multiple exhaustion, and the extreme item sampling plans. Table 7 compares single exhaustion to multiple exhaustion. In the case of tests with average biserial correlations of .80-.90, the ratios of SSEM's are 1.43, .74, .77, .97, and 1.44 for subtest sizes of 6, 10, 15, 20, and 30, respectively. Of these ratios, 3 out of 5 are less than 1.00 (viz., .74, .77, and .97), thus favoring single exhaustion item sampling. A total of 25 out of 35 ratios in Table 7 favor single exhaustion over multiple exhaustion. Table 8 compares single exhaustion to extreme item sampling. In the case of tests with average biserial correlations of .45-.95, the ratios of SSEM's are 1.73, 1.01, 1.58, .86, and .84 for subtest sizes of 6, 10, 15, 20, and 30, respectively. Of these ratios, 3 out of 5 are greater than 1.00 (viz., 1.73, 1.01, and 1.58), thus favoring the extreme plan over single exhaustion. A total of 19 out of 35 ratios in Table 8 favor the extreme plan over single exhaustion. Table 9 compares the extreme plan to multiple exhaustion item sampling. In the case of tests with average biserial correlations of .50-.70, the ratios of SSEM's are 1.17, 1.43, .84, 1.11, and .85 for subtest sizes of 6, 10, 15, 20, and 30, respectively. Of these

TABLE 7

RATIOS OF SUMS OF SQUARED ERRORS OF MEANS UNDER SINGLE AND MULTIPLE EXHAUSTION ITEM SAMPLING FOR DIFFERENT SUBTEST SIZES AND RANGES OF BISERIAL CORRELATION

Range of Biserial Correlation	Ratios of SSEM: Single/Multiple*				
	Subtest Size				
	6	10	15	20	30
.80-.90	1.43	.74	.77	.97	1.44
.45-.95	1.51	.83	1.28	1.19	.81
.50-.70	1.32	.49	.90	1.10	.93
.05-.95	.85	.95	.73	.87	.93
.20-.70	.85	1.02	.82	.99	.94
.05-.55	1.28	.65	.96	1.32	.92
.10-.20	.96	.90	.89	.73	.99

*Whenever the ratio is less than 1.00, single exhaustion item sampling is better than multiple exhaustion item sampling.

TABLE 6

RATIONS OF SUMS OF SQUARED ERRORS OF MEANS UNDER SINGLE EXHAUSTION AND EXTREME ITEM SAMPLING FOR DIFFERENT SUBTEST SIZES AND RANGES OF BISERIAL CORRELATION

Range of Biserial Correlations	Ratios of SSEM: Single/Extreme* Subtest Size				
	6	10	15	20	30
.80-.90	1.41	1.07	.79	.82	1.05
.45-.95	1.73	1.01	1.58	.86	.84
.50-.60	1.54	.71	.76	1.22	..80
.05-.95	.98	.97	.97	.96	.99
.20-.70	1.17	1.01	1.01	1.01	1.08
.05-.55	1.41	.83	.83	1.19	1.12
.10-.20	.97	1.01	1.12	.97	1.31

*Whenever the ratio is less than 1.00, single exhaustion item sampling is better than extreme item sampling.

TABLE 9

RATIOS OF SUMS OF SQUARED ERRORS OF MEANS UNDER MULTIPLE EXHAUSTION AND EXTREME ITEM SAMPLING FOR DIFFERENT SUBTEST SIZES AND RANGES OF BISERIAL CORRELATION

Range of Biserial Correlations	Ratios of SSEM: Multiple/Extreme* Subtest Size				
	6	10	15	20	30
.80-.90	.99	1.43	1.02	.85	.73
.45-.95	1.15	1.21	1.23	.73	1.03
.50-.70	1.17	1.43	.84	1.11	.86
.05-.95	1.16	1.02	1.34	1.10	1.06
.20-.70	1.37	.99	1.30	1.02	1.15
.05-.55	1.10	1.28	.75	.90	1.21
.10-.20	1.01	1.12	1.30	1.34	1.33

*Whenever the ratio is less than 1.00, multiple exhaustion item sampling is better than extreme item sampling.

ratios, 3 out of 5 are greater than 1.00 (viz., 1.17, 1.43, and 1.11), thus favoring extreme item sampling. A total of 26 out of 35 ratios in Table 9 favor the extreme plan over multiple exhaustion.

The results of estimating test population variance for the various plans are given in Tables 10, 11, and 12.

In considering the question of optimum number of people and items for estimating test population variance, it would seem that for tests with higher average biserial correlations, the optimal sampling plan would be one with small sized subtests. Thus, in Table 10, under the sampling plan 10/6/100 (6 items per subtest), tests with average biserial correlations of .80-.90 had a SSEV of 19399, whereas under the sampling plan 2/30/100 (30 items per subtest), the SSEV is 49603. When considering tests with average biserial correlations in the lower ranges of values (e.g., .10-.20, .05-.55), the optimal sampling plan consists of subtests with larger numbers of items per subtest. In fact, in the case of single exhaustion item sampling, the best plan involves a subtest size one half the size of the whole test. This can be seen in Table 10, where the smallest SSEV is 976. This value falls under the sampling plan 2/30/100 with 30 items per subtest, which is half the size of the whole test of 60 items. When considering tests with average biserial correlations in the middle ranges (e.g., .50-.60, .20-.70), the trend is not as clear, but sampling plans with small sized subtests do tend to be better. Thus, in Table 10 tests with aver-

TABLE 10

SUMS OF SQUARED ERRORS FOR VARIANCES UNDER TRADITIONAL AND SINGLE EXHAUSTION ITEM SAMPLING

Test Properties		Single Exhaustion Item Sampling Sums of Squared Errors					Traditional Sums of Squared Errors
Variance	Biserial Correlation	10/6/100	6/10/100	4/15/100 (T/k/n)[a]	3/20/100	2/30/100	1/60/100
326.87	.80-.90	19399	14695	26516	30188	49603	82237
228.89	.45-.95	13145	14676	17267	29097	37415	47463
142.15	.30-.60	9037	9209	13084	14527	16442	34203
135.66	.05-.95	10406	8651	7753	13060	14105	24451
115.18	.20-.70	7218	8613	7241	10266	8785	17005
48.28	.05-.55	6953	3970	3638	3476	2731	3942
23.05	.10-.20	4050	2929	1952	1397	976	1017

[a]Code: t = number of subtests, k = number of items per subtest, n = number of examinees per subtest

TABLE 11

SUMS OF SQUARED ERRORS FOR VARIANCES UNDER TRADITIONAL AND MULTIPLE EXHAUSTION ITEM SAMPLING

Test Properties		Multiple Exhaustion Item Sampling Sums of Squared Errors					Traditional Sums of Squared Errors
Variance	Biserial Correlation	20/6/50	20/10/30	20/15/20 (t/k/n)[a]	20/20/15	20/30/10	1/60/100
326.87	.80-.90	16375	24712	27290	30262	61943	82237
228.89	.45-.95	17164	15667	19790	26637	35876	47463
142.15	.50-.60	9334	8193	14093	15051	15724	34203
135.66	.05-.95	10350	8957	11608	12501	18975	24451
115.18	.20-.70	8785	9277	7928	12385	13975	17005
48.25	.05-.55	4537	5031	3322	4263	4286	3942
23.05	.10-.20	4350	2933	2161	1710	1156	1017

[a]Code: t = number of subtests, k = number of items per subtest, n = number of examinees per subtest

TABLE 12

SUMS OF SQUARED ERRORS FOR VARIANCES UNDER TRADITIONAL AND EXTREME ITEM SAMPLING

Test Properties		Extreme Item Sampling Sums of Squared Errors					Traditional Sums of Squared Errors
Variance	Biserial Correlation	500/6/2	300/10/2	200/15/2 (t/k/n)[a]	150/20/2	100/30/2	1/60/100
326.87	.80-.90	45163	55241	106734	104023	152011	82237
228.89	.45-.95	28468	37949	42869	75183	88135	47463
142.15	.50-.60	19527	24962	38051	26478	32405	34203
135.66	.05-.95	13583	25093	20994	18246	26296	24451
115.18	.20-.70	20105	16684	17457	23933	22920	17005
48.28	.05-.55	10414	6943	7109	5211	9722	3942
23.05	.10-.20	7750	4431	3006	2852	2411	1017

[a]Code: t = number of subtests, k = number of items per subtest, n = number of examinees per subtest

TABLE 13

RATIOS OF SUMS OF SQUARED ERRORS OF VARIANCES UNDER SINGLE AND MULTIPLE EXHAUSTION ITEM SAMPLING FOR DIFFERENT SUBTEST SIZES AND RANGES OF BISERIAL CORRELATION

Range of Biserial Correlations	Ratios of SSEV: Single/Multiple*				
	Subtest Size				
	6	10	15	20	30
.80-.90	1.18	.59	.97	.99	.80
.45-.95	.77	.94	.87	.79	1.04
.50-.60	.97	1.12	.93	.97	1.05
.05-.95	1.01	.97	.67	1.04	.74
.20-.70	.82	.93	.91	.83	.63
.05-.55	1.33	.79	1.09	.81	.64
.10-.20	.93	.99	.89	.82	.84

*Whenever the ratio is less than 1.00, single exhaustion item sampling is better than multiple exhaustion item sampling.

TABLE 14

RATIOS OF SUMS OF SQUARED ERRORS OF VARIANCES UNDER SINGLE EXHAUSTION AND EXTREME ITEM SAMPLING FOR DIFFERENT SUBTEST SIZES AND RANGES OF BISERIAL CORRELATION

Range of Biserial Correlations	Ratios of SSEV: Single/Extreme* Subtest Size				
	6	10	15	20	30
.80-.90	.43	.27	.25	.29	.33
.45-.95	.46	.39	.40	.28	.42
.50-.60	.46	.37	.39	.55	.51
.05-.95	.77	.34	.37	.71	.54
.20-.70	.36	.52	.41	.43	.38
.05-.55	.58	.57	.51	.67	.28
.10-.20	.52	.66	.64	.49	.40

*Whenever the ratio is less than 1.00, single exhaustion item sampling is better than extreme item sampling.

TABLE 15
RATIOS OF SUMS OF SQUARED ERRORS OF VARIANCES UNDER MULTIPLE EXHAUSTION AND EXTREME ITEM SAMPLING FOR DIFFERENT SUBTEST SIZES AND RANGES OF BISERIAL CORRELATION

Range of Biserial Correlations	Ratios of SSEV: Multiple/Extreme*				
	Subtest Size				
	6	10	15	20	30
.80-.90	.36	.45	.25	.29	.41
.45-.95	.60	.41	.46	.35	.41
.50-.60	.48	.33	.43	.37	.49
.05-.95	.76	.36	.55	.69	.72
.20-.70	.44	.56	.45	.52	.61
.05-.55	.43	.72	.47	.82	.44
.10-.20	.56	.66	.72	.60	.48

*Whenever the ratio is less than 1.00, multiple exhaustion item sampling is better than extreme item sampling.

age biserial correlations of .50-.60 under sampling plan 10/6/100 (k=6) had a SSEV of 9037, whereas under sampling plan 2/30/100 (k=30), the SSEV is 16442.

An examination of the data of Tables 13, 14, and 15 clearly indicates that for estimating population variance, single exhaustion item sampling is superior to either multiple exhaustion or the extreme item sampling. This can be seen by the preponderance of ratios less than 1.00. A total of 27 out of 35 ratios in Table 13 favor single exhaustion over multiple exhaustion and a total of 25 out of 35 ratios in Table 14 favor single exhaustion over the extreme plan.

SUMMARY

The purpose of this study was to help determine optimum item sampling plans for use in school evaluation projects. The findings presented provided the basis for the following conclusion.

How does the range of biserial correlations between item response and ability, i.e., item discrimination, affect the choice of optimum item sampling plan?

When estimating test population mean, no matter what the values of the biserial correlations are, better estimates are given by the item sampling plans with fewer items and more subtests. Similar conclusions were drawn by Barcikowski (1970). Shoemaker (1971) also recommended the use of a large number of subtests. The results of this study support their findings.

The case is different when estimating test population variance. Here, the optimum sampling plan depends very much on the biserial correlation of the test in question. With biserial correlations in the higher ranges, e.g., .80-.90 and .45-.95, employment of subtests with fewer items produce better estimates. In the case of tests with biserial correlations in the lower ranges, e.g., .10-.20 and .05-.55, just the opposite is true. Here, subtests containing more items produced better variance estimates. Specifically, in the case of single subtests one half the size of the whole test. For tests with average biserial correlations in the middle ranges, e.g., 150-.60 and .20-.70, the results, although not as clear-cut, also tended to favor smaller subtests. These conclusions support the findings in Barcikowski's study of an interaction effect between item characteristics (biserial correlations) and sampling plan. They, however, contradict Shoemaker's conclusion that item-examinee sampling plans have, for all practical purposes, the same standard error in estimating population variance. This study indicates that such is not the case when item biserial correlations are considered.

Should single or multiple exhaustion of the set of items on a test be employed? The results favor single exhaustion item sampling for the estimation of population mean and variance for most practical purposes. This is decidedly true in the case of population variance estimates. In the case of population mean estimates, extreme item sampling had a slight edge over single exhaustion item sampling, but the difference is not sufficient to warrant the production of the extremely large number of subtests

necessary to implement the extreme item sampling design.

RECOMMENDATIONS

For educational practitioners who desire guidelines in the application of item sampling techniques to school evaluation situations, the results of this study suggest the following recommendations:

1. When the parameter of primary importance is the population mean test score, small sized subtests should be used (e.g., subtests of 6 items, with the data from this study), and as many of these subtests as cost factors indicate to be practicable.

2. When the parameter of primary importance is the population test variance, attention should be given to the item discrimination of the test. With tests of low item discrimination (e.g., with biserial correlations in the range .05-.35), larger subtests should be used, specifically, subtests approaching half the size of the whole test. In all other instances, smaller subtests may be used.

3. Single exhaustion of the item set is sufficient for either population mean or variance estimates.

REFERENCES

Barcikowski, R.S. Optimum use of the item sampling technique in obtaining test norms. Unpublished doctoral dissertation, State University of New York at Buffalo, 1970.

Lord, F.M. and Novick, M.R. Statistical Theories of Mental Test Scores. Reading, Massachusetts: Addison-Wesley, 1968.

Shoemaker, D.M. Further results on the standard errors of estimate associated with item-examinee sampling procedures. Journal of Educational Measurement, 1971, 8, 215-220.